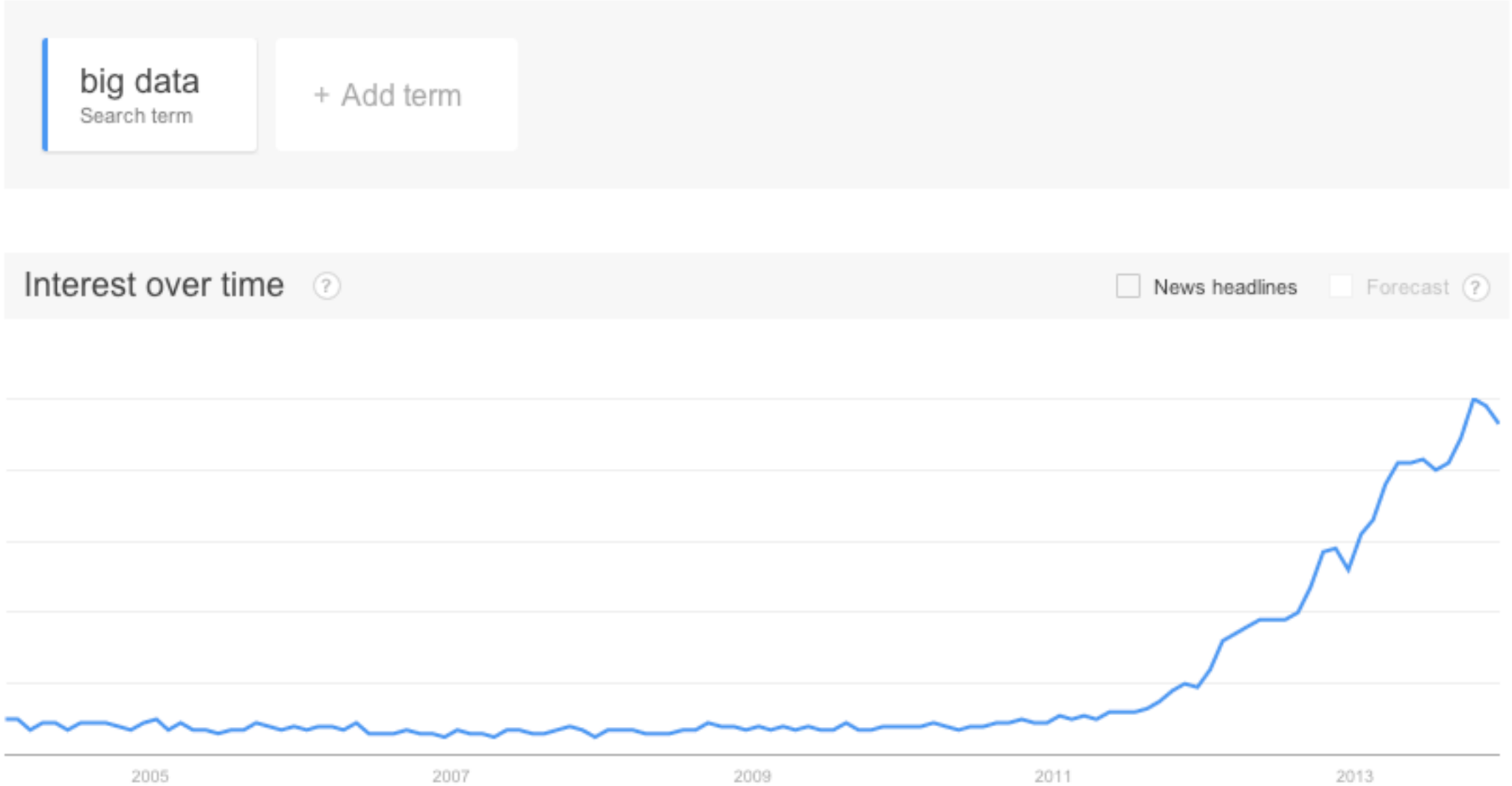
Defining Big Data

Matthew A Levin MD
Assistant Professor
Department of Anesthesiology
Icahn School of Medicine at Mount Sinai

Mount Sinai

Big Data is hot



http://www.google.com/trends/explore#q=big%20data&cmpt=q

Credit to Prof. Osmar Zaïane of the University of Alberta for slide concept

What data is “big data”?

- financial
- retail
- environmental
- social (twitter feeds, facebook posts)
- ...all data is becoming “big data”

WHAT IS BIG DATA?

VOLUME

Large amounts of data.

VELOCITY

Needs to be analyzed **quickly.**

VARIETY

Different types of structured and unstructured data.

Key questions enterprises are asking about Big Data:

How to store and protect big data?

How to backup and restore big data?

How to organize and catalog the data that you have backed up?

How to keep costs low while ensuring that all the critical data is available when you need it?

WHAT ARE THE VOLUMES OF DATA THAT WE ARE SEEING TODAY?

30 billion pieces of content were added to Facebook this past month by 600 million plus users.

zynga

Zynga processes 1 petabyte of content for players every day; a volume of data that is unmatched in the social game industry.

YouTube

More than 2 billion videos were watched on YouTube... yesterday.

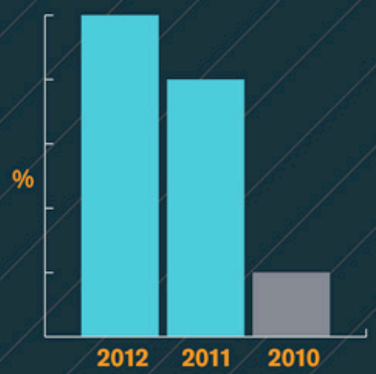
LOL!

The average teenager sends **4,762 text messages** per month.

32 billion searches were performed last month... on Twitter.

Source: Gartner

Everyday business and consumer life creates 2.5 quintillion bytes of data per day.



90% of the data in the world today has been created in the last two years alone.

Source: IBM

WHAT DOES THE FUTURE LOOK LIKE?

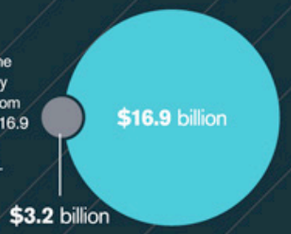
Worldwide IP traffic will **quadruple by 2015.**

By 2015, nearly **3 billion people**

will be online, pushing the data created and shared to nearly **8 zettabytes.**

HOW IS THE MARKET FOR BIG DATA SOLUTIONS EVOLVING?

A new IDC study says the market for big technology and services will grow from $3.2 billion in 2010 to $16.9 billion in 2015. **That's a growth of 40%** CAGR.



58% of respondents expect their companies to increase spending on server backup solutions and other big data-related initiatives within the next three years.

Source: Economist Business Unit

2/3rds of surveyed businesses in North America said big data will become a concern for them within the next five years.

Source: Economist Business Unit

Asigra.

http://www.asigra.com/blog/big-data-infographic-what-big-data

www.wipro.com

BIG DATA

Big Data is data that is too large, complex and dynamic for any conventional data tools to capture, store, manage and analyze.

The right use of Big Data allows analysts to spot trends and gives niche insights that help create value and innovation much faster than conventional methods.

The "three V's", i.e the Volume, Variety and Velocity of the data coming in is what creates the challenge.

VOLUME

Amount of Big Data stored across the world (in petabytes)

VARIETY

PEOPLE TO PEOPLE
NETIZENS, VIRTUAL COMMUNITIES, SOCIAL NETWORKS, WEB LOGS...

PEOPLE TO MACHINE
ARCHIVES, MEDICAL DEVICES, DIGITAL TV, E-COMMERCE, SMART CARDS, BANK CARDS, COMPUTERS, MOBILES...

MACHINE TO MACHINE
SENSORS, GPS DEVICES, BAR CODE SCANNERS, SURVEILLANCE CAMERAS, SCIENTIFIC RESEARCH...

VELOCITY

2.9 MILLION
EMAILS SENT EVERY SECOND

20 HOURS
OF VIDEO UPLOADED EVERY MIN

50 MILLION
TWEETS PER DAY

CASE STUDY - Healthcare

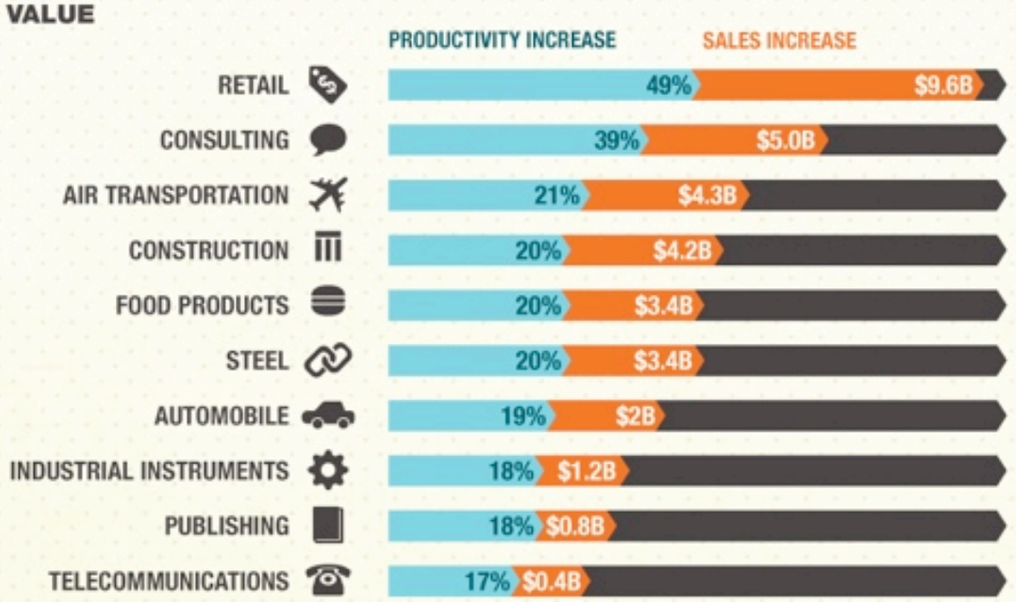
$300 billion is the potential annual value to Healthcare

57.6% OF ORGANIZATIONS SURVEYED SAY THAT BIG DATA IS A CHALLENGE

72.7% CONSIDER DRIVING OPERATIONAL EFFICIENCIES TO BE THE BIGGEST BENEFIT OF A BIG DATA STRATEGY

50% SAY THAT BIG DATA HELPS IN BETTER MEETING CONSUMER DEMAND AND FACILITATING GROWTH

VALUE



40% PROJECTED GROWTH IN GLOBAL DATA CREATED PER YEAR

5% PROJECTED GROWTH IN GLOBAL IT SPENDING PER YEAR

The estimated size of the digital universe in 2011 was 1.8 zettabytes. It is predicted that between 2009 and 2020, this will grow 44 fold to 35 zettabytes per year. A well defined data management strategy is essential to successfully utilize Big Data.

Sources - 1 Reaping the Rewards of Big Data - Wipro Report 2 Big Data: The Next Frontier for Innovation, Competition and Productivity - McKinsey Global Institute Report 3 comScore, Radicati Group 4 Measuring the Business Impacts of Effective Data - study by University of Texas, Austin 5 US Department of Labour.

DO BUSINESS BETTER

NYSE:WIT | OVER 130,000 EMPLOYEES | 54 COUNTRIES | CONSULTING | SYSTEM INTEGRATION | OUTSOURCING

WIPRO Applying Thought

Visualization: Seventhinc

http://www.alphasixcorp.com/images/big-data-infograph.jpg

Defining Big Data

- The three V's
  - **Volume** – the amount of data is large, easily terabytes daily
  - **Velocity** – data comes at high speed; time-sensitive analysis must be fast
  - **Variety** – data is heterogeneous

Laney D. *3D Data Management: Controlling Data Volume, Velocity, and Variety*. META group Inc., 2001. http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf. *Accessed on Dec 13, 2013.*

Defining Big Data

- Others have added additional V's:
  - **Veracity** – sources must be trusted, and results of analysis must be trusted
  - **Value** – big data only is only of value if it is of high **Veracity** with low **Vulnerability**

Defining Big Data

- A buzzword for the new infrastructures able to deal with large volumes of heterogeneous data
- Mostly based on
  - Distributed file systems
  - Distributed processing
- The relationship to data mining and analytics is evolving

Credit to Prof. João Gama of the University of Porto Portugal for slide concept

Does Anesthesiology really have big data? - Volume

- Potentially - yes (raw physio data, aggregate nationwide data)
- Practically (for most sites currently) - no

Volume - example

- HR - for a 120 min case
  - BPM every 5 min: 24 data points
  - BPM every 15 sec: 480 data points
  - EKG sampled at 250 Hz: 30,000 data points
- Big data? Not really

Volume - another way

- Single case: 1 MB
- 200 cases/day: 200 MB
- 50,000 cases/year: 50 GB
- 51.4* million cases/year in US: ~51 terabytes (49.01 if you're counting)
- Big data? Getting closer

http://www.cdc.gov/nchs/fastats/insurg.htm

How many cases does AQI have?*

- 13 million cases to date
- BUT - Only 1 million have AIMS data
- So, thats about 1 TB of data - maybe

*2013 data http://www.aqihq.org/

Does Anesthesiology really have big data? - Velocity

- Intraoperative data is real time
- Hundreds of thousands of cases daily natiowide
- Real time analysis of intraoperative trends *has the potential* to alter management and improve outcomes
- **However -**
  - No one has yet shown that it *can* or *does* either of these things

Does Anesthesiology really have big data? - Variety

- A variety of data types
  - Physiologic data - discrete, continuous
  - Demographic data - text, categorical
  - Medication data
  - Event data (timestamps)
  - Imaging data (glidescope,TEE...)

Does Anesthesiology really have big data? - Variety

- A variety of data sources
  - AIMS*
  - EMR*
  - Billing databases
  - Closed claims database
  - CMS data
- From a variety of places
  - Academic medical centers
  - Community hospitals
  - Ambulatory surgical centers

Summary

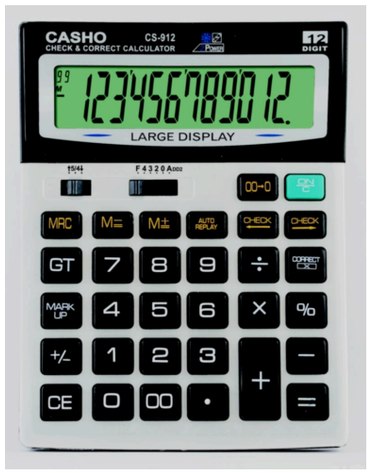
- We *definitely* have Variety
- We *probably* have Velocity
- We *kind of* have Volume

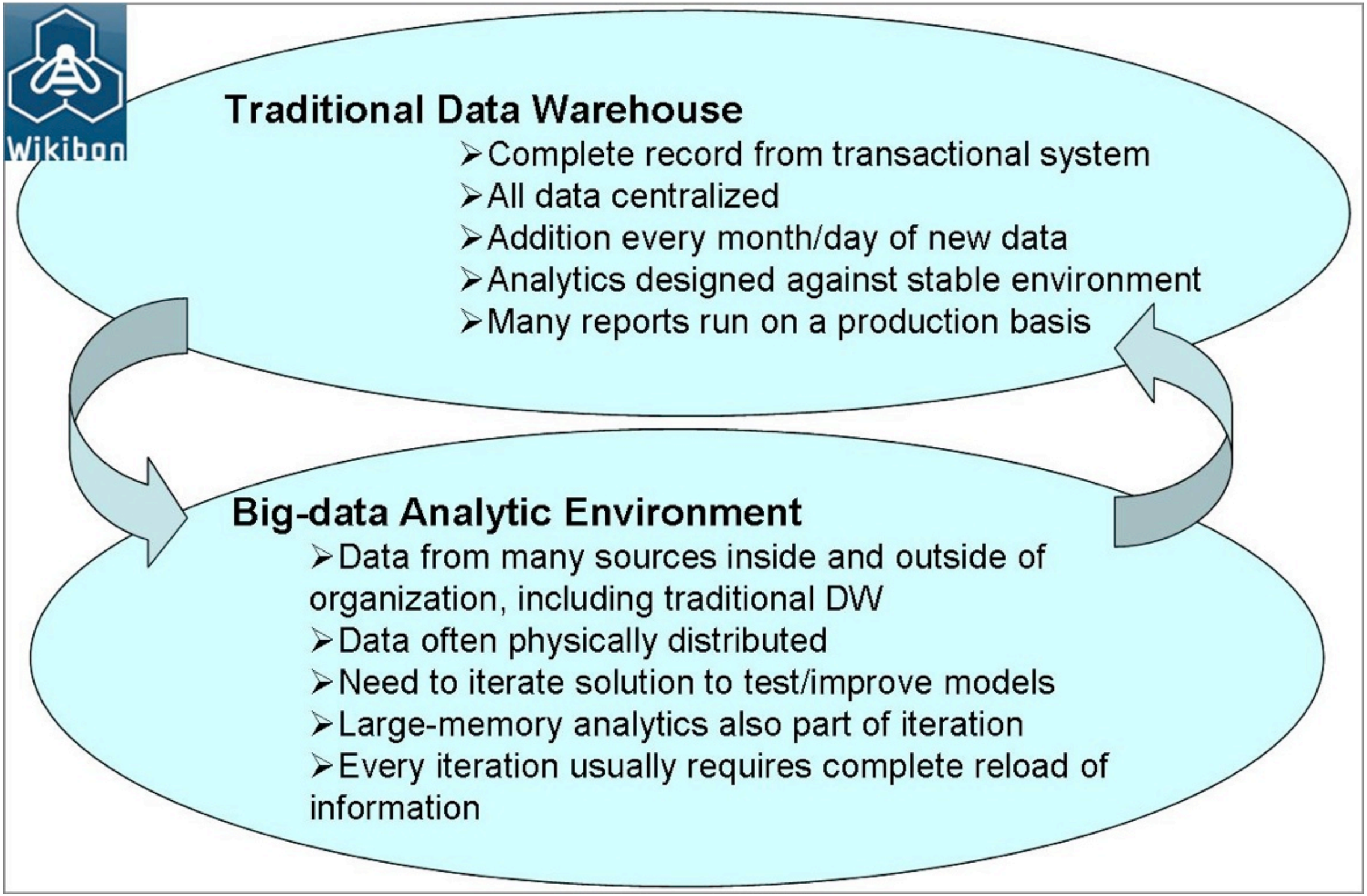
Limitations of traditional analytics

Questions

- True incidence of...
- True cause of...
- Does technology X really allow us to...
- Are we really compliant with...

Traditional Tools





http://wikibon.org/wiki/v/Enterprise_Big-data

Dirty secret

- Not all “Big Data” requires new analytic techniques or new infrastructure
- Machines are faster
- Storage is cheap and plentiful
- “older” technologies like relational databases and SQL still work just fine

The real dirty secret

Many big data techniques are designed to pre-process large datasets so they can be imported,analyzed using and manipulated using “traditional” tools

New technologies

- “NoSQL”
  - everything is a blob of data
  - no tables and relations, just key-value pairs
  - do not guarantee atomicity, consistency, isolation or durability (ACID principles) BUT
    - fast for retrieval and appending
- Examples
  - Memcached - web server caching
  - MongoDB - document store (content management)
  - MUMPS* - the engine that drives Epic

http://en.wikipedia.org/wiki/NoSQL

New technologies

- Map-Reduce
  - “Map” chunks of data to nodes
  - “Reduce” each chunk to an answer, reduce all answers to a single final answer
    - storage and computation distributed across multiple nodes
    - fault tolerant, redundant
- Examples
  - Google’s original implementation
  - Hadoop (Cloudera)

http://en.wikipedia.org/wiki/Map_reduce

Hadoop

- Originally written for a open source web search engine to
- Backed by Yahoo since 2006, “web scale” since 2008
- Java based
- Batch - **not** realtime
- Requires programming team - not turnkey or end-user friendly

Common theme of new tools

- Even more so than velocity or variety, really designed to handle *Volume*
- Require dedicated teams to implement and use

Conclusion

- Anesthesia has lots of data but its not yet Big Data on the scale of FB, Twitter
- It may never be on that scale
- Traditional analytic techniques and tools will work fine for quite a while
- All we need to do is use them!

(But hey, using buzzwords is good too)

Thank you!

Mount Sinai